Statistics and Management of WET Tests

Washington State has implemented a whole effluent toxicity (WET) program using narrative criteria. Our WET regulation was praised by cities, industries, and environmentalists. Compliance with narrative criteria is determined by comparing response at an effluent concentration representing the point of compliance to a control using single comparison hypothesis testing. Single comparisons allow direct control of both false positive and false negative error rates. Discarding tests without good concentration-response relationships further reduces false positives. This regulation was inspired by EPA's *Technical Support Document for Water Quality-based Toxics Control* (TSD) and a similar approach in North Carolina validated by instream bioassessment (See Eagleson, et al. 1990. *E.T.&C.* 9:1019-1028.). The TSD focuses on numeric criteria, but current EPA regulations allow either numeric or narrative criteria.

EPA's water quality criteria have acute and chronic points of compliance. The chronic point of compliance is the edge of a mixing zone where receiving water must be suitable for long-term habitation. Inside of the mixing zone and close to the discharge point is the acute point of compliance where there must be no lethality. The area between these points of compliance need not be fit for long-term habitation but must not be lethal to passing organisms. WET tests only assessing survival apply at the acute point of compliance. Short-term chronic and critical life stage tests apply at the chronic point of compliance to assess suitability for long-term habitation by aquatic species.

Measuring compliance with narrative WET criteria is equivalent to the numeric criteria in the TSD. The TSD proposes numeric criteria based on toxic units. Acute toxic units (ATU) are determined by dividing the LC_{50} into 100. The acute toxicity detection limit is LC_{50} = 100% effluent (1.0 ATU). The TSD sets the acute numeric criterion at 0.3 ATU to reflect the typical ratio of an LC_1 to an LC_{50} and be more protective. This criterion is less than the detection limit of 1.0 ATU. When dilution at the point of compliance is insufficient to translate the criterion of 0.3 ATU to a WET limit above the self-imposed detection limit, the solution is to change the criterion of 0.3 ATU to a criterion of 1.0 ATU calculated by dividing the NOEC into 100. Two acute criteria with two different calculations are unnecessarily complicated since a criterion of 1.0 ATU calculated by dividing the NOEC into 100 will always work.

The TSD proposes a chronic numeric criterion of 1.0 chronic toxic units (CTU) calculated by dividing the IC_{25} into 100. EPA chronic testing manuals (section 9.7.2.3) question the reliability of linear interpolation to determine the IC_{25} . The EC_{25} and other effect levels have been proposed for calculating CTU from quantal measurements such as bivalve development. Probit cannot be run with less than 2 partial responses which happens frequently because choosing a concentration series for effluents is guesswork. When insufficient partial mortalities or other circumstances prevent the use of Probit, Spearman-Karber provides only the 50% effect level establishing another self-imposed detection limit.

The TSD bases the validity of the 25% effect level for regulating chronic toxicity on its equivalence to the NOEC. If comparability to the NOEC is the standard for validity, it is simpler to use the NOEC than attempt to establish comparability of effect levels. The TSD offers the alternative of dividing the NOEC into 100 for calculating CTU. As explained above, using the NOEC is also the only method for calculating ATU under all circumstances. Acute and chronic criteria are both 1.0 toxic units if NOECs are used.

Further reasoning shows that toxic units and NOECs are unnecessary. Dropping toxic units simply means that the NOEC must be ≥ the concentration of effluent at the point of compliance. Simplifying further, the same concept says there must be no statistically significant toxicity at the point of compliance, which is how compliance with narrative criteria is determined.

Point estimates are best for comparing test results, but cannot separate real effects within a single test from effects due to variability in measurements. Point estimates have confidence intervals which extend in both directions requiring judgment on which side to err to take them into account. A confidence interval does not assess variability in the measurements used in determining concentration means involved in deriving point estimates.

Hypothesis testing allows control of both false positives and false negatives. Each hypothesis test takes into account variance across replicates before determining differences are statistically significant. *Alpha* can be chosen to minimize compliance failures that are not due to toxicity. *Alpha* approximates the false positive rate when values being compared are close. As the difference between values being compared increases, the false positive rate decreases without intervention, and the false negative rate becomes a concern. The false negative rate exceeds the false positive rate because the null hypothesis is no toxic effect. Burden of proof is on the demonstration of toxicity. To minimize undetected toxicity, the number of replicates can be increased.

Because multiple comparisons control the experiment-wise error rate, false negatives increase as the number of concentrations increases. Fortunately, it is not necessary to know the no effects concentration to determine compliance with narrative criteria. Only when mixing zones are unknown, are NOECs useful. When the effluent concentration at the point of compliance is known, multiple comparisons only reduce statistical sensitivity compared to single comparisons.

Single comparisons allow efficient use of test results. Many tests with three replicates don't meet assumptions for valid parametric multiple comparison. Nonparametric alternatives have no critical values for three replicates. If the number of concentrations is increased or if *alpha* is decreased, critical values for nonparametric multiple comparisons can be found lacking for four, five, or more replicates. Nonparametric single comparisons have critical values for three replicates and above.

Even though single comparisons are used, tests are usually conducted with five concentrations to evaluate concentration-response. Identifying anomalous tests reduces false positives. An increasing concentration-response indicates toxicity as opposed to other stresses. Method variability or lab error rarely produce a good concentration-response relationship.